Linear Regression (Revisited)

STAT 245

Sept 5, 2024

Data

We will consider a small dataset from an article by J.S. Martin and colleagues, titled *Facial width-to-height ratio is associated with agonistic and affiliative dominance in bonobos (**Pan paniscus**)*

```
bonobos <-
  read_csv(file='http://sldr.netlify.app/data/bonobo_faces.csv')
```

Facial Width-Height Ratio

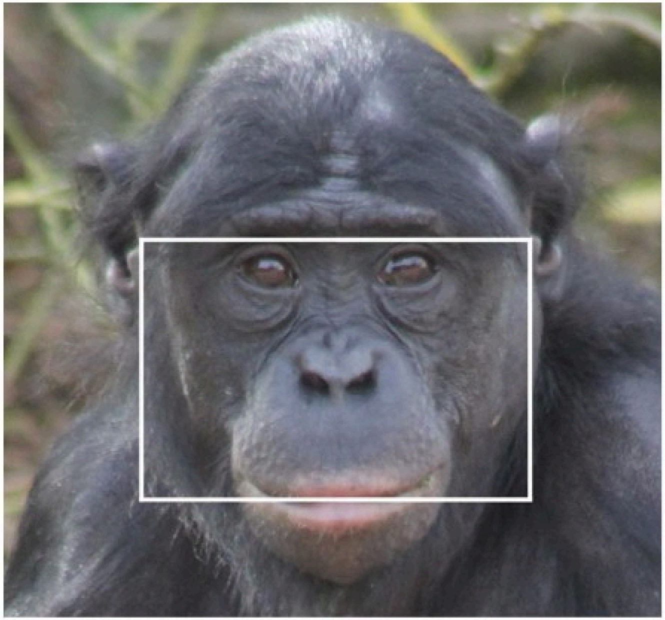


Figure 1. Facial width-to-height ratio. fWHR was measured by dividing the bizygomatic breadth (white box width) by the distance between the brow ridge and upper lip (white box height). The tragus was used for reference when facial hair covered the maximal cheek prominence.

Data: Variables

We *will* explore data - let's think a bit first.

- `Name` of the individual
- `Group`: zoo where the individual lives
- `Sex`, "Male or "Female"
- `Age` in years
- `fWHR`, facial width-height ratio
- `AssR`, assertiveness score
- `normDS`, another dominance score
- `weight` in kilograms

Simple Linear Regression

("Simple" means "one predictor")

$$y = \ ?$$

Choosing Response

- Response is the variable of greatest interest
- Response is what you may want to *predict*
- Response may be causally dependent on predictor

Choosing Predictor(s)

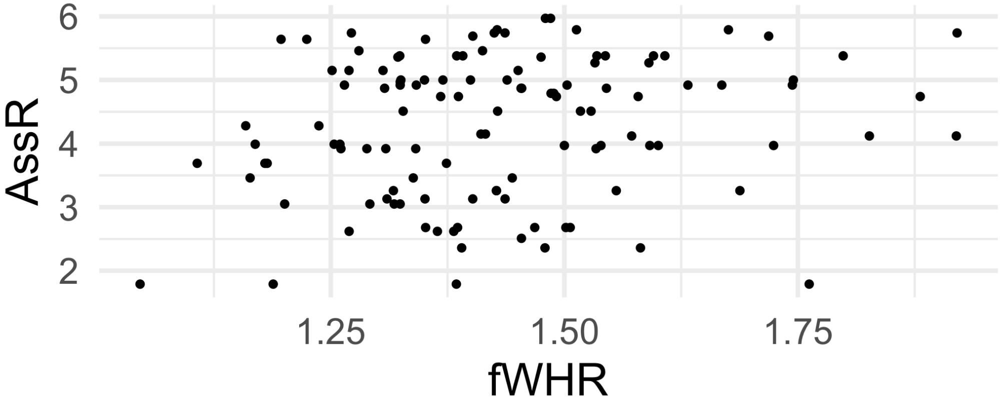
- Response may be harder to measure, and predictor easier
- (Should have data on predictor)
- **Choosing predictor that "looks best" in data -> biased inference**

For Bonobos:

- Response
- Predictor

Plan. STOP! Explore.

```
gf_point(AssR ~ fWHR, data = bonobos)
```



To fit model: `lm()`

```
m1_simple <- lm(AssR ~ fWHR,
                data = bonobos)
```

Print Results

Just Parameter Estimates

```
coef(m1_simple)
```

```
## (Intercept)        fWHR
##    2.547350    1.213902
```

Print Results

Full Summary

```
summary(m1_simple)
```

Summary

```
##
## Call:
## lm(formula = AssR ~ fWHR, data = bonobos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8963 -0.8405  0.2307  0.8442  1.6485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5474     0.8518   2.991  0.00341 **
## fWHR          1.2139     0.5907   2.055  0.04213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 115 degrees of freedom
## Multiple R-squared:  0.03542,     Adjusted R-squared:  0.02704
## F-statistic: 4.223 on 1 and 115 DF,  p-value: 0.04213
```

(Part of) Regression Equation

$$y = \_\_ + \_\_x \ldots$$

But the Line isn't Perfect

There is some error ϵ in our model for the data, so we should keep track of it.

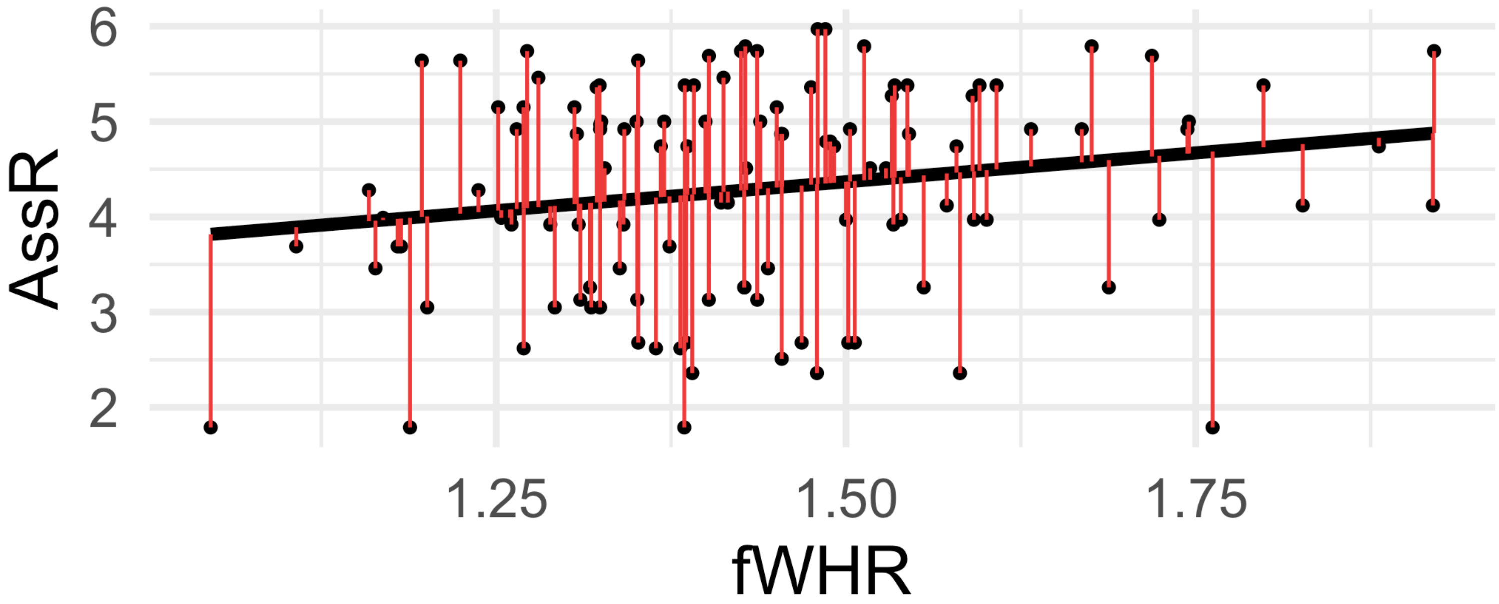
$$y = \beta_0 + \beta_1 x + \epsilon$$

ϵ is NOT *one number*

The amount of error is different for (almost) every data point!

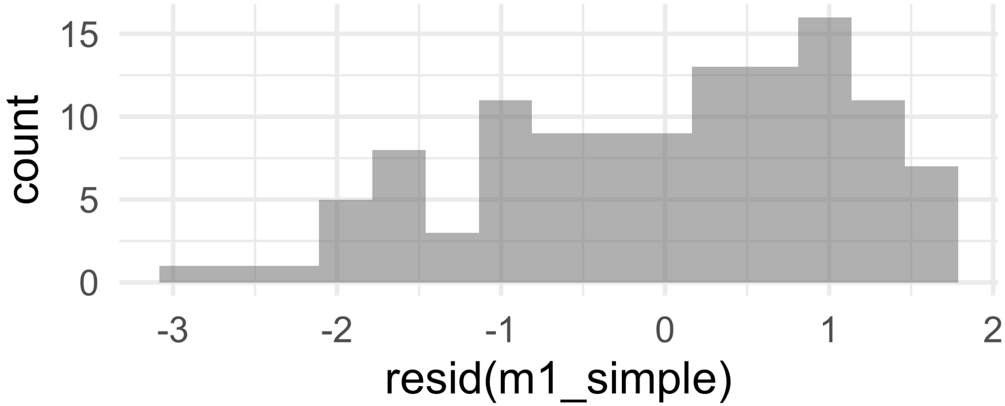
- How can we express this mathematically?

Regression Residuals = "errors"



Summarize: Distribution?

```
gf_histogram(~resid(m1_simple),
             bins = 15)
```



Back to Summary

```
##
## Call:
## lm(formula = AssR ~ fWHR, data = bonobos)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -2.8963 -0.8405  0.2307  0.8442  1.6485
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)   2.5474     0.8518   2.991  0.00341 **
## fWHR          1.2139     0.5907   2.055  0.04213 *
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 1.09 on 115 degrees of freedom
## Multiple R-squared:  0.03542,    Adjusted R-squared:  0.02704
## F-statistic: 4.223 on 1 and 115 DF,  p-value: 0.04213
```

Complete Regression Equation

$$y = \beta_0 + \beta_1 x + \epsilon,$$

where

$$\epsilon \sim N(0, \sigma)$$

Bonobo Regression Equation

$$y = \underline{\qquad\quad} + \underline{\qquad\quad}\, x + \epsilon,$$

$$\text{where } \epsilon \sim N(0, \underline{\qquad\quad})$$